The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data

Colin Batchelor[1], Thomas Bittner[2], Karen Eilbeck[3], Chris Mungall[4], Jane Richardson[5]; Rob Knight[6]; Jesse Stombaugh[6], Craig Zirbel[7], Eric Westhof[8], and Neocles Leontis[7]
[1]Royal Society of Chemistry, Cambridge, UK CB4 0WF; [2] Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA; [3]Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112, USA; [4]Life Sciences Division, Lawrence Berkeley National Lab, CA 94720, USA; [5]Department of Biochemistry, Duke University Medical School, Durham, NC 27710, USA; [6]Department of Chemistry and Biochemistry, University of Colorado at Boulder, Boulder, CO 80309, USA; [7]Departments of Chemistry and of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43402, USA; [8]Architecture et réactivité de l'ARN, Université Louis Pasteur de Strasbourg, Institut de Biologie Moléculaire et Cellulaire du CNRS, F-67084 Strasbourg, France.

Abstract

Biomedical Ontologies are intended to integrate diverse biomedical data to enable intelligent data-mining and facilitate translation of basic research into useful clinical knowledge. We present the first version of RNAO, an ontology for integrating RNA 3D structural, biochemical and sequence data. While each 3D data file depicts the structure of a specific molecule, such data have broader significance as representatives of classes of homologous molecules, which, while differing in sequence, generally share core structural features of functional importance. Thus, 3D structure data gain value by being linked to homologous sequences in genomic data and databases of sequence alignments. Likewise genomic data can increase in value by annotation of shared structural features, especially when these can be linked to specific functions. The RNAO is being developed in line with the developing standards of the Open Biomedical Ontologies (OBO) Consortium.

Introduction

The aim of the RNA Ontology Consortium (ROC)[1] is "to create an integrated conceptual framework—an RNA ontology—with a common, dynamic, controlled and structured vocabulary to describe and characterize RNA sequences, secondary structures, three-dimensional structures and dynamics pertaining to RNA function." Other kinds of experiment that are useful to RNA biochemists and bioinformaticists include chemical probes and thermodynamic measurements. Previous work in this field includes the RiboWeb ontology,[2] which was part of a knowledge base for studying the bacterial ribosome, the Multiple Alignment Ontology for nucleic acid and protein sequences[3] and RNAML,[4] which is an actively-used XML schema for exchanging information about RNA secondary structures, tertiary structures, sequences and sequence alignments. The immediate context of the RNA Ontology is the Open Biomedical Ontologies (OBO) project,[5] which seeks to coordinate the development of biomedical ontologies. Small molecules are dealt with by ChEBI,[6] macromolecular sequences (DNA, RNA and protein) by the Sequence Ontology[7] and proteins by the Protein Ontology.[8] The RNAO is distinct from its neighbors but will share relationships and refer to terms from the other ontologies where necessary.

We set out the paper as follows: we briefly describe the chemical structure of the RNA molecule and then describe how to represent (1) base pairing and other pairwise interactions, (2) motifs and (3) backbone conformations based on the hierarchical nature of RNA structure. We will also describe the relationship to the Sequence Ontology. The RNAO is developed using Protégé as an OWL[1] ontology and is also available in OBO format. We illustrate what can be done within the limitations of OWL; however a full treatment of RNA structure requires first-order logic. RNAO is freely available.[2]

RNA

Ribonucleic acid (RNA) molecules consist of nucleotide (nt) units, which themselves consist of heterocyclic nucleobases covalently bonded to ribose rings which are connected covalently to the ribose rings of other nucleotides through phosphate groups.

[1] http://www.w3.org/TR/owl-features/
[2] http://code.google.com/p/rnao

The combination of base and ribose is called a nucleoside. Each nucleoside has three interacting edges, the Watson-Crick edge, the Hoogsteen edge and the sugar edge as shown in Fig. 1. These edges are sets of hydrogen-bond donors and hydrogen-bond acceptors located on the same stretch of the boundary of the nucleoside. They are illustrated for adenosine in Fig. 1. The nucleotide units themselves are linked one to the next in a directional manner, usually by connection of the 3' position of a nucleotide to the 5' position of the next nucleotide in the chain via the phosphate group (see Fig. 1).

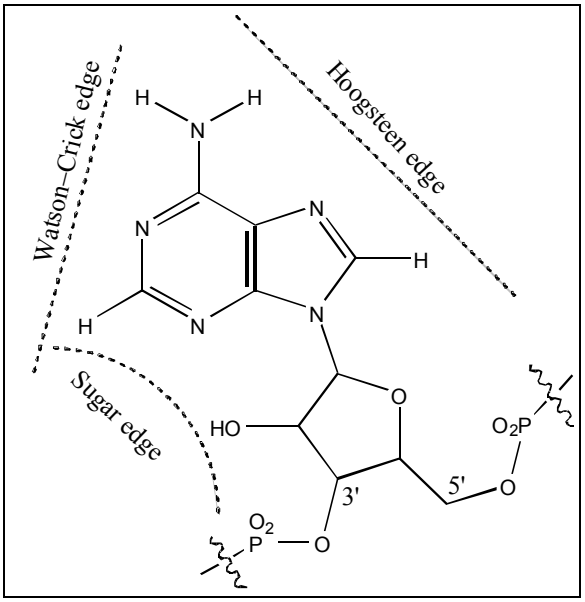


Figure 1. The Watson–Crick, Hoogsteen and Sugar edges on an adenosine nucleotide.

We follow Villanueva-Rosales and Dumontier,[9] who base their ontology on atoms and bonds, but we modify their approach by treating the nucleotides as the objects and the interactions between them as the relations. Thus we have two fundamental relations, the covalently_bonded_to relation, and the pairs_with relation.

The folding of the RNA chain brings together pairs of short sequence segments that are Watson-Crick complementary to form anti-parallel double helices consisting of stacked Watson-Crick basepairs. Helices are the simplest and most regular RNA 3D motifs. The set of Watson-Crick paired helices comprise the secondary structure of the RNA. Some RNA molecules can form more than one secondary structure and can be induced by appropriate perturbations to switch between them. The looping of the chain forms other motifs called hairpin loops, many of which are structured by specific sets of interactions, including base-pairing and base-stacking and often, base-phosphate interactions. Segments of sequence joining two helices can also form structured motifs called internal loops. Finally, multi-helix junction loops result when three or more helical segments are joined together. Junction loops provide branch points in RNA molecules. RNA 3D motifs recur in numerous RNA molecules encoded by genes from different families in very different organisms. Recurrent 3D motifs often play similar roles in different RNA molecules. For example, junction loops provide branch points, kink-turn internal loops provide flexible hinges and GNRA hairpin loops mediate tertiary interactions. Motifs combine to define characteristic RNA folds or domains.

Base pairing

We start with the basepair classification proposed by Leontis and Westhof,[10] which places RNA basepairs in distinct, geometrically defined classes that are mutually exhaustive and disjoint. The pairwise interactions are hydrogen bonds between atoms in adjacent nucleosides, and as such we define the interactions in terms of *edges* (see Fig. 1). To a first approximation:

(1) *each edge of a nucleoside may interact only with a single edge of a different nucleoside*

Because OWL can only handle binary relations, we have to specialize the pairs_with relation for each combination of interacting edges. With six different combinations of edge interaction (WC-WC, H-H, S-S, WC-H, WC-S and H-S), and two relative orientations (*cis* and *trans*) for the interaction of the nucleosides, there result twelve basepairing classes in the Leontis-Westhof scheme and eighteen base pairing relations as shown in Table 1. We can express statement (1) formally by declaring each of these relations to be *disjoint* from other relations, which means for example that if X pairs_with_CWH Y then there is no Z such that X pairs_with_CWW Z. The logical definition for a family 1 base pair is written:

family_1_base_pair = hasPart some (Nucleobase and pairs_with_CWW some Nucleobase)

in OWL Manchester syntax,[11] and this is sufficient for a reasoner to classify a base pair with the correct pairing relation into the correct LW family.

Motifs

By specializing the covalently_bonded_to relation and pairs_with relation it is possible to create rudimentary definitions of most motifs, and it is straightforward to generate RNAO-specific first-order logic representations of a given RNA structure from a plain text file. However, because all but the

very simplest motifs contain cyclically-connected nucleotides, and OWL cannot handle cycles, it is impossible for this part of the ontology to be represented in OWL in such a way that reasoners can deal with it.

Further, it is possible that some motifs will be best described by formal definitions, whereas other more complex motifs may be best described by statistical or machine learning approaches.

Backbone conformers

The backbone in RNA molecules is a chain of covalently-bonded atoms which are parts either of the phosphate group (O5', P, O3') or of the ribose rings (C3'-C4'-C5'). We are interested in RNA backbone conformations for two reasons: (1) particular RNA motifs can also be described as a sequence of backbone conformers, and (2) they provide sites for catalysis or interaction with ions, proteins, small molecules, proteins, and other nucleic acids or segments of the same RNA.

We are using the ROC backbone committee's 2-character notation[12] for the conformations of suites, which are the stretches of backbone between two ribose rings. Each of their 54 suite conformers is a cluster of datapoints in the 7-dimensional space of the backbone dihedral angles. Suites and nucleotides provide alternative ways to partition the RNA molecule, but we are exploring whether the ontology can simply treat suite conformers as qualities of the covalent connection between nucleotides.

RNAO and the Sequence Ontology

The Sequence Ontology (SO) is a structured controlled vocabulary for the description of biological sequence. SO is used by model organism genome communities for the annotation of genomic sequence and will provide the basic terms to describe sequence features for RNAO. SO will be extended to provide terms to describe discontiguous regions. This will be necessary to describe many secondary and tertiary structural motifs. SO also includes a number of RNA motif terms that will be transitioned to RNAO.

Conclusions

We have presented a rudimentary version of RNAO which contains logical definitions that can be used by a reasoner to classify base pairings into the twelve categories of Leontis and Westhof and outlined how to incorporate 3D motifs and backbone configurations into the ontology. We have also shown what can be done in OWL for interoperability with other OBO ontologies and what needs to be represented in first-order logic.

Acknowledgements

The ROC is supported by a Research Coordination Network (RCN) grant from the National Science Foundation (grant no. 0443508). SO is supported by the NHGRI, via the Gene Ontology Consortium, HG004341.

References

1. Leontis NB *et al.* The RNA Ontology Consortium: an open invitation to the RNA community. RNA, 2006;12;533
2. Altman R *et al.* RiboWeb: An ontology-based system for collaborative molecular biology. IEEE Intelligent Systems 1999;14(5);68-76.
3. Thomson JD *et al.* MAO: A Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Research 2005;33(13);4164-4171.
4. Waugh A *et al.* RNAML: a standard syntax for exchanging RNA information. RNA, 2002;8;707.
5. Smith B *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 2007;25; 1251.
6. Degtyarenko K *et al.* ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res. 2008;36;D344.
7. Eilbeck K *et al.* The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6(5):R44.
8. Natale DA *et al.* BMC Bioinformatics 2007;8(Suppl 9);S1.
9. Villanueva-Rosales N and Dumontier M. 2007. Describing chemical functional groups in OWL-DL for the classification of chemical compounds, in OWL: Experiences and Directions (OWLED 2007), co-located with European Semantic Web Conference (ESWC2007), Innsbruck, Austria.
10. N. B. Leontis and E. Westhof. 2001. Geometric nomenclature and classification of RNA base pairs, RNA, 2001;7;499-512.
11. M. Horridge *et al.* 2006. The Manchester OWL Syntax, in OWL Experiences and Directions Workshop, 2006.
12. Richardson JS *et al.* RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution), RNA, 2008;14;465-481.

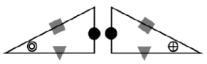
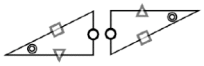
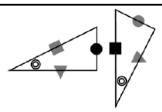
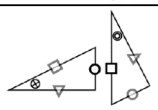
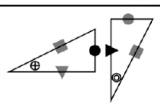
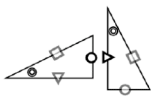
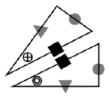
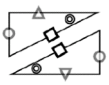
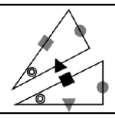
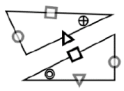
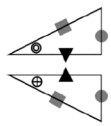
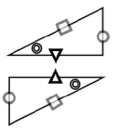
	Relations					
Classes	**Base pairing relations**		**Inverse of**	**Disjoint with other**	**Symbol**	**Triangle Abstraction**
family_1_base_pair	pairs_with_WW	pairs_with_CWW	symmetric	W-first pairings W-second pairings		
family_2_base_pair		pairs_with_TWW	symmetric	W-first pairings W-second pairings		
family_3_base_pair	pairs_with_WH	pairs_with_CWH	pairs_with_CHW	W-first pairings H-second pairings		
	pairs_with_HW	pairs_with_CHW	pairs_with_CWH	H-first pairings W-second pairings		
family_4_base_pair	pairs_with_WH	pairs_with_TWH	pairs_with_THW	W-first pairings H-second pairings		
	pairs_with_HW	pairs_with_THW	pairs_with_TWH	H-first pairings W-second pairings		
family_5_base_pair	pairs_with_WS	pairs_with_CWS	pairs_with_CSW	W-first pairings S-second pairings		
	pairs_with_SW	pairs_with_CSW	pairs_with_CWS	S-first pairings W-second pairings		
family_6_base_pair	pairs_with_WS	pairs_with_TWS	pairs_with_TSW	W-first pairings S-second pairings		
	pairs_with_SW	pairs_with_TSW	pairs_with_TWS	S-first pairings W-second pairings		
family_7_base_pair	pairs_with_HH	pairs_with_CHH	symmetric	H-first pairings H-second pairings		
family_8_base_pair		pairs_with_THH	symmetric	H-first pairings H-second pairings		
family_9_base_pair	pairs_with_HS	pairs_with_CHS	pairs_with_CSH	H-first pairings S-second pairings		
	pairs_with_SH	pairs_with_CSH	pairs_with_CHS	S-first pairings H-second pairings		
family_10_base_pair	pairs_with_HS	pairs_with_THS	pairs_with_TSH	H-first pairings S-second pairings		
	pairs_with_SH	pairs_with_TSH	pairs_with_THS	S-first pairings H-second pairings		
family_11_base_pair	pairs_with_SS	pairs_with_CSS	symmetric	S-first pairings S-second pairings (this will change in v2)		
family_12_base_pair		pairs_with_TSS	symmetric	S-first pairings S-second pairings (this will change in v2)		

Table 1: Base pairing relations in RNAO.